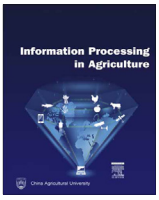
Available at www.sciencedirect.com

INFORMATION PROCESSING IN AGRICULTURE 1 (2014) 42–50

journal homepage: www.elsevier.com/locate/inpa



Time-series prediction of shellfish farm closure: A comparison of alternatives☆

Ashfaqur Rahman *, *Md Sumon Shahriar, Claire D'Este, Greg Smith, John McCulloch, Greg Timms*

Autonomous Systems Program, CSIRO Computational Informatics, Hobart, TAS 7000, Australia

ARTICLE INFO

Article history:
Received 3 September 2013
Received in revised form
13 May 2014
Accepted 19 May 2014
Available online 24 May 2014

Keywords:
Aquaculture
Shellfish farm closure
Data mining
Time series data

ABSTRACT

Shellfish farms are closed for harvest when microbial pollutants are present. Such pollutants are typically present in rainfall runoff from various land uses in catchments. Experts currently use a number of observable parameters (river flow, rainfall, salinity) as proxies to determine when to close farms. We have proposed using the short term historical rainfall data as a time-series prediction problem where we aim to predict the closure of shellfish farms based only on rainfall. Time-series event prediction consists of two steps: (i) feature extraction, and (ii) prediction. A number of data mining challenges exist for these scenarios: (i) which feature extraction method best captures the rainfall pattern over successive days that leads to opening or closure of the farms?, (ii) The farm closure events occur infrequently and this leads to a class imbalance problem; the question is what is the best way to deal with this problem? In this paper we have analysed and compared different combinations of balancing methods (under-sampling and over-sampling), feature extraction methods (cluster profile, curve fitting, Fourier Transform, Piecewise Aggregate Approximation, and Wavelet Transform) and learning algorithms (neural network, support vector machine, k-nearest neighbour, decision tree, and Bayesian Network) to predict closure events accurately considering the above data mining challenges. We have identified the best combination of techniques to accurately predict shellfish farm closure from rainfall, given the above data mining challenges.

* Corresponding author. Tel.: +61 3 6232 5536; fax: +61 3 6232 5050.
E-mail address: Ashfaqur.rahman@csiro.au (A. Rahman).

☆ This work was supported in part by a grant from Tasmanian Government which is administered by the Tasmanian Department of Economic Development, Tourism and the Arts and in part by the CSIRO Food Futures Flagship.
Peer review under the responsibility of China Agricultural University.

Production and hosting by Elsevier

1. Introduction

Consumption of contaminated shellfish can pose severe health hazards to humans and may even cause death. Authorities such as the Tasmanian Shellfish Quality Assurance Program (TSQAP) are responsible for monitoring the safety of product coming from commercial shellfish farms. These organizations have the authority to close the farms if they are concerned about water quality at the growing site.

Contamination can be categorised into biotoxin, chemical, or microbial groups. Biotoxins are due to naturally occurring

http://dx.doi.org/10.1016/j.inpa.2014.05.001

algal blooms, chemical issues are generally from incidents or spills in the waterway, and microbial pollutants are from thermotolerant coliforms (commonly from water runoff from agricultural land use). This work focuses on microbial contamination of the waterway that typically occurs when fresh water enters the waterway following rainfall. Given historical records of daily rainfall data and closure status of shellfish farms, the research presented in this paper aims to develop a supervised learning framework to predict closure of shellfish farms from time-series rainfall patterns. Data mining/machine learning techniques (e.g. [1,2]) are rarely applied to aquaculture problems; with one exception being the prediction of harmful algal blooms [3]. Our Aquaculture Decision Support (AquaDS) project has previously investigated class imbalance issues [4], dealing with missing sensor values [5], problems related to relocating models to locations where we do not have sufficient closure examples [6], and identifying causes of closure [7]. However time series framework is yet to be explored.

The time-series prediction framework used in this paper is presented in Fig. 1. The historical rainfall data is divided in windows and representative features are extracted from each window. It is expected that features leading to closure are different from features when the farm is open. During the training phase, a closure status ('Open'/'Close') is associated with features computed from each window. A classifier is then trained on the features of the balanced data set. During real-time prediction of events, features are computed from a window of recent rainfall data. The features are then fed to the trained classifier that produces the predicted closure status.

For most sites in Tasmania closures are a rare event, leading to a class imbalance. A classifier trained on imbalanced data [8,9] is likely to predict the minority class (in this case 'Close') with low accuracy. This leads to the question: what is the best method to deal with the class imbalance problem? Time-series events involve feature extraction. It is not clearly known which feature extraction method best captures the rainfall pattern over successive days that lead to opening or closure of the farms. The research and results presented in this paper aim to identify the answers to these questions.

The performance of the time series prediction framework in general depends on a number of things: (a) class balancing method, (b) the feature extraction method, and (c) the classifier. In this paper we aim to identify the best combination of class balancing method, feature set, and classifier that can predict shellfish farm closure with high accuracy. We have considered both under-sampling and over-sampling class balancing methods. Features were extracted using the following feature extraction methods: cluster profile, curve fitting, Fourier transform, Piecewise Aggregate Approximation (PAA), and Wavelet transform. We have used the following classifiers: neural network, support vector machine, k-nearest neighbour, decision tree, and Bayesian Network. These steps are detailed in the following sections. We have compared their performance, identified the best possible combination, and suggested reasons for the results under the light of the above-mentioned data mining challenges.

2. Class imbalance

Class imbalance [26] refers to the scenario where the number of samples available for a particular class is significantly higher than that for other classes. The class with a high number of samples is called the 'majority' class whereas the other is called 'minority' class. The shellfish farm closure events occur with low frequency and thus the farms remain open most of the time. This results in the 'Open' class outnumbering the 'Close' class in the supervised classification framework. We have analysed two alternative approaches (under-sampling and over-sampling [25,31,32]) to deal with the class imbalance problem.

2.1. *Under-sampling*

In the random under-sampling method, the data from the majority class is sampled down to match the number of samples in the minority class. Classifiers are then trained on the balanced data set. The sampling process is conducted randomly. Although commonly used, a criticism of this method is that the sampled down data do not always follow the original distribution of the original data [10]. This sometimes leads to poor classification performance.

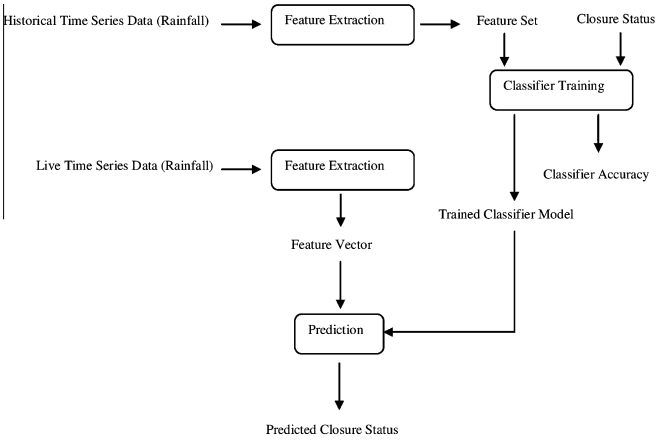


Fig. 1 – Time series prediction framework for shellfish farm closure based on rainfall.

2.2. Over-sampling

The random over-sampling approach increases the number of samples of the minority class to match that of the majority class by randomly sampling the data in the minority class with replacement. Classifiers are trained on the balanced data set. This is also a commonly used approach and the distribution of the original data set is preserved.

3. Feature extraction algorithms

We have considered six feature extraction methods from the time series data: Sample as is, cluster profile [11], curve fitting [12,33], Fourier Transform [13], Piecewise Aggregate Approximation [14], and Wavelet Transform [13]. The feature extraction methods are explained in the following sections. We assume that n successive samples in time series x are represented as $(x_1, x_2, \ldots, x_n)$.

3.1. Sample as is

In this particular method, the actual samples are used as features. The time series is presented by a total of n features $(x_1, x_2, \ldots, x_n)$.

3.2. Cluster profile features

In this method n samples in the time series $(x_1, x_2, \ldots, x_n)$ are clustered into n_c clusters. The clustering process groups the samples of identical values into similar clusters. The feature vector is composed of the cluster labels of the samples. Let the cluster label for sample x_i be $c(x_i)$ where $1 \leqslant i \leqslant n$. The feature vector is represented by $(c(x_1), c(x_2), \ldots, c(x_n))$.

3.3. Curve fitting features

In this method the time series is approximated by a polynomial and the parameters of the polynomial are used as features. Given a polynomial order n_p, the polynomial is expressed as:

$$x_t = a_0 + a_1 \times t + a_2 \times t^2 + \cdots + a_{n_p} \times t^{n_p} \quad (1)$$

where t represents the time stamp. The best fit parameters $(\hat{a}_0, \hat{a}_1, \cdots, \hat{a}_{n_p})$ are computed from the time series data $(x_1, x_2, \ldots, x_n)$. The time series is represented by a total of $n_p + 1$ features and the feature vector is $(\hat{a}_0, \hat{a}_1, \cdots, \hat{a}_{n_p})$.

3.4. Fourier transformation features

In this method the frequency domain representation of the time series is obtained using a discrete Fourier Transform (DFT) [29,30]. Given the finite list of equally-spaced successive time series samples $x_1, x_2, \ldots, x_n$ DFT produces a list of n coefficients $f_1, f_2, \ldots, f_n$ of a finite combination of complex sinusoids, ordered by their frequencies. The magnitude of a coefficient $|f_i|$ is used as feature where $1 \leqslant i \leqslant n$. The time series is represented by a total of n features and the feature vector is $(|f_1|, |f_2|, \ldots, |f_n|)$.

3.5. Piecewise Aggregate Approximation (PAA) features

In this method the local structure of a time series is captured by aggregating samples over time. Given a block size b, the time series of length n is divided into n/b equal-sized blocks. The mean value of data points falling in each block is calculated. Let m_j be the mean value of block j where $1 \leqslant j \leqslant n/b$. The vector of these mean values $(m_1, m_2, \ldots, m_{n/b})$ becomes the feature vector and a total of n/b features is thus computed. The time series is normalized to have mean 0 and standard deviation 1 before converting it to PAA.

3.6. Wavelet features

This is also a frequency domain representation of the time series obtained using a Discrete Wavelet Transform (DWT) [27,28,30]. The process starts by pairing up successive pairs in the time series and the summation of these pairs represent the series in the next level. This process is repeated recursively, pairing up the sums to provide the next scale and it produces a tree-like structure. The values in the non-leaf nodes of the tree are used as Wavelet features. Given a list of $n = 2^m$ time series samples a total of $n + \frac{n}{2} + \frac{n}{2^2} + \cdots + 1$ Wavelet coefficients represented the feature vector.

4. Learning algorithms

We have used a total of five different classifiers to evaluate the effectiveness of the features: neural network, support vector machine, k-nearest neighbour, decision tree, and Bayesian Network. Each of these classifiers is briefly presented in the following sections.

4.1. Neural network

A neural network [15] is made up of a number of interconnected processing nodes. Each node processes information as a function i.e. by generating dynamic responses to external inputs. Neural networks are arranged in layers. Layers are composed of a number of nodes and each node contains an activation function. Patterns are presented to the network via the input layer. The output from the input layer is communicated to the hidden layers. The actual processing is done within the hidden layers via a system of weighted connections. The hidden layers connect to an output layer where the answer is provided as output. A learning rule recursively modifies the weights of the connections according to the input patterns and output targets. We have used two types of neural networks in the experiments: the Multi Layer Perceptron (MLP) and Radial Basis Function (RBF) networks.

4.2. Support vector machine

A support vector machine or SVM [16] transforms the data into a higher dimension using a kernel function and finds the best linear hyperplane that separates the patterns of one class from those of the other class. The best hyperplane for an SVM refers to the one with the maximum margin between the classes. The support vectors are the data points that are closest to the separating hyperplane. These points

are on the boundary of the slab. In out experiments we have used Radial Basis Function (RBF) kernel.

4.3. *k-Nearest neighbour*

In *k*-Nearest neighbour (*k*-NN) classification the distance between a test pattern and all the patterns in the training set is computed. The distance can be calculated using Euclidian distance or Manhattan distance. The probable classes receive a vote from each of the *k* patterns that are closest to the test pattern in terms of distance. The class that obtains the highest vote is considered to be the class of the test pattern.

4.4. *Decision tree*

A decision tree (DT) builds classification models that take the form of a tree structure. It breaks down a dataset into smaller subsets recursively and at the same time maintains an incompletely built decision tree. The end result is a tree with intermediate/decision nodes and terminal/classification nodes. A decision node offers branching opportunities. A terminal node offers a classification verdict. The root decision node in a tree corresponds to the best attribute in terms of classification capability. Both categorical and numerical data can be handled by decision trees. The selection of a decision node while building the decision tree, is guided by information gain, Gini index [17] etc.

4.5. *Bayesian Network*

Bayesian Networks (BN) represents the classification framework as probabilistic graph model [18]. Each node in the graph represents a random variable and the edges between the nodes represent probabilistic dependencies (or causal relationship) among the corresponding random variables. These conditional dependencies in the graph are estimated based on historical data. The Bayesian Network can be constructed using automatic discovery algorithms or from domain knowledge. Given a pattern, the probability of a particular class is expressed as a product of prior and conditional probabilities. The factoring of the product is done following the structure of the Bayesian Network. BN operates on discrete features. All continuous attributes are discretized for using in the Bayesian Network.

5. Results and discussions

We have conducted a set of experiments to identify the best combination of balancing method, feature extraction method and classifier to accurately predict shellfish farm closures based solely on rainfall data. As mentioned in Section 1 rainfall is the prime cause of microbial-based farm closures in Tasmania as concluded in previous research [19]. We have collected time series rainfall data and farm closure status on six different shellfish farms in Tasmania: Big Bay Zone B, Big Bay Zone C, Duck Bay, Dunnalley Bay Zone A, Hastings Bay, and Montagu. Rainfall data is obtained from SILO [20] and Bureau of Meteorology sensors [21]. We have expressed farm closure prediction as a classification problem where features extracted (from a time window of fourteen days) form the input and the closure decision (open/close) represents the class.

Table 1 – Information on data gathered and used in the research.

Location	# Instances	Class distribution %	
		Open	Close
Big Bay Zone B	4756	71.30	28.70
Big Bay Zone C	4756	72.69	27.31
Duck Bay	4789	72.35	27.65
Dunnalley Bay A	3589	88.44	11.56
Hastings Bay	2137	78.43	21.57
Montagu	2767	61.47	38.53

The information on data gathered from the sensors at different locations is presented in Table 1. The time series rainfall and closure pattern for each location is presented in Fig. 2. Note that in all the data sets the percentage of the 'Close' class is smaller than that of the 'Open' class. This leads to a class imbalance problem. We have dealt with the imbalance problem using under-sampling, and over-sampling. We utilized the WEKA [22] implementation of different classifiers and default parameter settings of these classifiers are used in the experiments. We have evaluated a total of six classification algorithms and the parameter settings of each classifier are presented in Table 2.

We have used a time window of fourteen days to extract all the features (except wavelet). The input feature vector thus has a length of fourteen for all features except PAA (explained next). Data was partitioned into ten clusters to generate cluster profile based features. A polynomial curve of order seven was fitted to the time series window to generate the curve fitting features. Magnitudes of the frequency domain were used as features in the Fourier Transform method. A block size of two was used to generate PAA features and the length of the feature vector was seven. The window size for the wavelet transform needs to be a power of two and we have used a window size of sixteen to compute wavelet features. Wavelet coefficients were computed using complex-valued Daubechies' wavelets. The coefficients with a normalised information gain ratio of at least 0.01 were considered and a total of seventy-five wavelet coefficients were used as features. All the features in the feature vector are considered to have equal weights [23,24].

The results presented here are discussed in three sections. The first section analyses the performance of different balancing algorithms. The second section presents and analyses performance of different features using different classifiers. Finally the third section identifies the best combination. In the first two sections, results are compared based on Matthews Correlation Coefficient (MCC). Given true positive rate, TP (percentage of correctly classified instances of True or Closure class), true negative rate, TN (percentage of correctly classified instances of False or Open class), false positive rate, FP (percentage of instances classified as True but actually False), and false negative rate FN (percentage of instances classified as False but actually True), the MCC is computed as

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (2)$$

MCC obtains a maximum score (+1) when both True Positive and True Negative are 100% and a minimum score (−1) when False Positive and False Negative are 100%. MCC is a

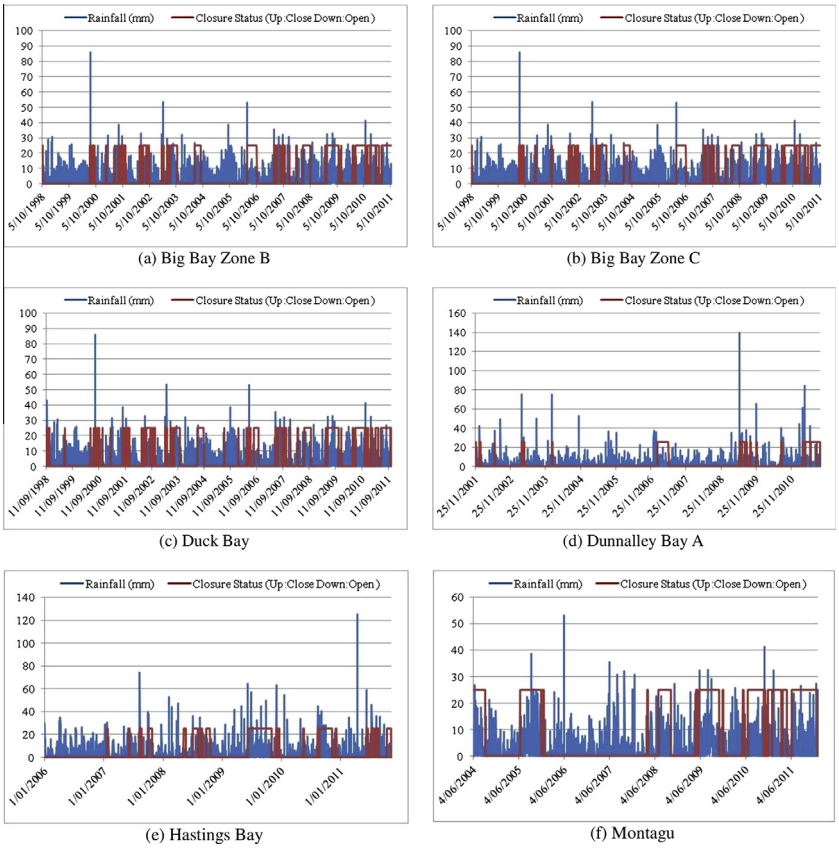


Fig. 2 – Rainfall and closure patterns at different shellfish farms used in this study.

Table 2 – Configuration of the different classification algorithms.

Classification algorithm	Parameter setting
Decision tree	Algorithm: C4 Confidence factor: 0.25
Support vector machine	Algorithm: SMO c: 1.0 Epsilon: 10^{-12} Kernel: Polynomial
Multi Layer Perceptron (MLP)	Hidden Layers: (No. of features)/2 Learning Rate: 0.3 Momentum: 0.2 Epochs: 500
Bayesian Network	Estimator algorithm: simple estimator Alpha (initial count): 0.5 Search algorithm: K_2
RBF	Minimum standard deviation: 0.1 Number of clusters: 2 Ridge: 10^{-8}
k-NN classifier	k = 1

good measure to deal with imbalance as it will penalize results that undermine the minority class. The accuracy of detection is presented in the third section.

5.1. *Implication of sampling*

The MCC scores obtained under three scenarios are presented in Fig. 3: no balancing, balancing using over-sampling, and balancing using under-sampling. All the scores are averaged across all the six classifiers. Note that both balancing algorithms improve the MCC score in almost all locations. This is because of the improvement in classification accuracy of the 'Close' class after balancing. The improvement with balancing is relatively smaller in Montagu. This is because the percentage of the 'Close' class is relatively higher in Montagu than other regions (Table 1). On a head-to-head performance comparison, over-sampling and under-sampling are almost equally capable. The over-sampling process, however, preserves the underlying distribution of the majority class. We present results based on over-sampling only for the rest of the paper without loss of generality.

5.2. *Implication of feature set*

In this section we analyse the performance (MCC score) of different feature extraction methods. The performance of the feature sets in each location using different classifiers is presented in Table 3 to Table 8.

The rainfall is used as the feature using the *sample as is* method (Table 3). In the first three locations the Bayesian Network is the best performer whereas in the last three locations the SVM is the best performer. The Bayesian Network

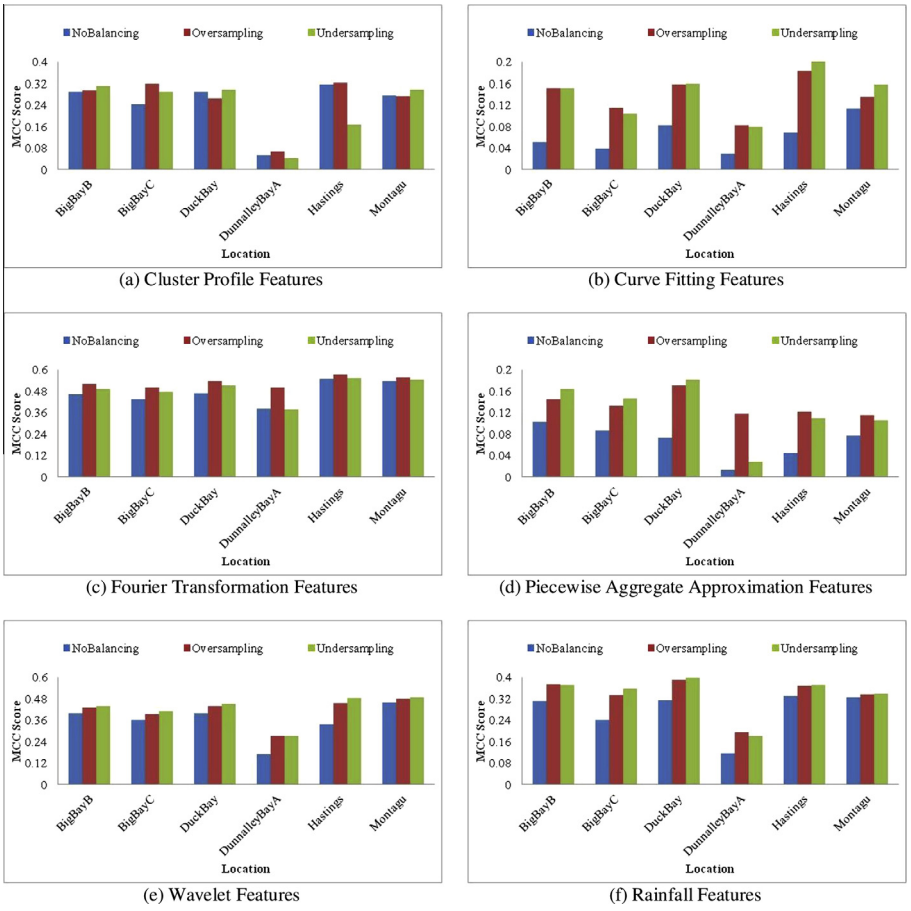


Fig. 3 – Performance comparison of different balancing methods.

achieves the highest average MCC score (0.436) and the *k*-NN classifier achieves the lowest average MCC score (0.182) across all locations. Rainfall data can be noisy and as raw features are used "as is" in this method, classifiers like *k*-NN are expected to suffer as they do not transform the feature space. SVM transforms the feature space to a higher dimension and the Bayesian Network discretizes the features and this leads to thesuperior performance with this method.

While using *Cluster Profile Features* (Table 4) the Bayesian Network is the best performer the majority of the time. This is because cluster labels are discrete and no additional discretization step is required in the Bayesian Network. The performance of *Curve Fitting features* (Table 5) is relatively inferior to other features. This is because a polynomial of a certain order fails to capture all the windows.

The *Fourier Transform Features* (Table 6) performs best with the *k*-NN classifier. The MCC scores are relatively higher than for other features. The reason is that the transitions are presented by high frequency components whereas steady states are represented by low frequency components. This can be explained with an example from Fig. 4 where two segments of the time series data from Big Bay B is presented. In Fig. 4(a) transition in states occurs due to change in rainfall pattern whereas no transition between states occurs in Fig. 4(b). The FFT response of these rainfall patterns are shown next to the rainfall plots. It can be observed that higher frequency components in Fig. 4(a) are much stronger than that in Fig. 4(b). The ratio of low and high frequency components in Fig. 2(a and b) are 0.366 and 0.488, respectively. This indicates that low frequency components are stronger in

Table 3 – MCC scores obtained using Rainfall Features. Bold indicates the best performance in a row.

	DT	SVM	MLP	BN	RBF	k-NN
Big Bay B	0.332	0.429	0.420	**0.519**	0.331	0.207
Big Bay C	0.272	0.395	0.355	**0.491**	0.322	0.163
Duck Bay	0.314	0.461	0.432	**0.542**	0.361	0.229
Dunnalley Bay A	0.075	**0.294**	0.236	0.187	0.258	0.114
Hastings	0.227	**0.485**	0.437	0.453	0.388	0.215
Montagu	0.226	**0.43**	0.399	0.423	0.376	0.169

Table 4 – MCC scores obtained using Cluster Profile Features. Bold indicates the best performance in a row.

	DT	SVM	MLP	BN	RBF	k-NN
Big Bay B	0.198	0.388	0.342	**0.424**	0.374	0.037
Big Bay C	0.218	**0.415**	0.349	0.396	0.355	0.179
Duck Bay	0.232	0.218	0.296	**0.441**	0.369	0.04
Dunnalley Bay A	0.041	0.038	0.073	**0.21**	0.065	−0.019
Hastings	0.242	0.404	0.288	**0.471**	0.363	0.167
Montagu	0.219	0.334	0.252	**0.371**	0.303	0.155

Table 5 – MCC scores obtained using Curve Fitting Features. Bold indicates the best performance in a row.

	DT	SVM	MLP	BN	RBF	k-NN
Big Bay B	0.161	0.129	**0.212**	0.162	0.17	0.079
Big Bay C	0.162	0.031	**0.164**	0.138	0.135	0.058
Duck Bay	0.201	0.109	**0.214**	0.153	0.179	0.098
Dunnalley Bay A	0.147	−0.055	0	0.177	**0.236**	−0.012
Hastings	0.215	0.011	0.225	0.225	**0.258**	0.167
Montagu	**0.23**	0.049	0	0.199	0.201	0.128

Table 6 – MCC scores obtained using Fourier Transformation Features. Bold indicates the best performance in a row.

	DT	SVM	MLP	BN	RBF	k-NN
Big Bay B	0.647	0.461	0.503	0.357	0.333	**0.814**
Big Bay C	0.612	0.441	0.488	0.324	0.315	**0.818**
Duck Bay	0.638	0.465	0.548	0.379	0.381	**0.806**
Dunnalley Bay A	0.675	0.319	0.405	0.621	0.243	**0.723**
Hastings	0.69	0.5	0.605	0.428	0.371	**0.838**
Montagu	0.637	0.458	0.569	0.414	0.407	**0.856**

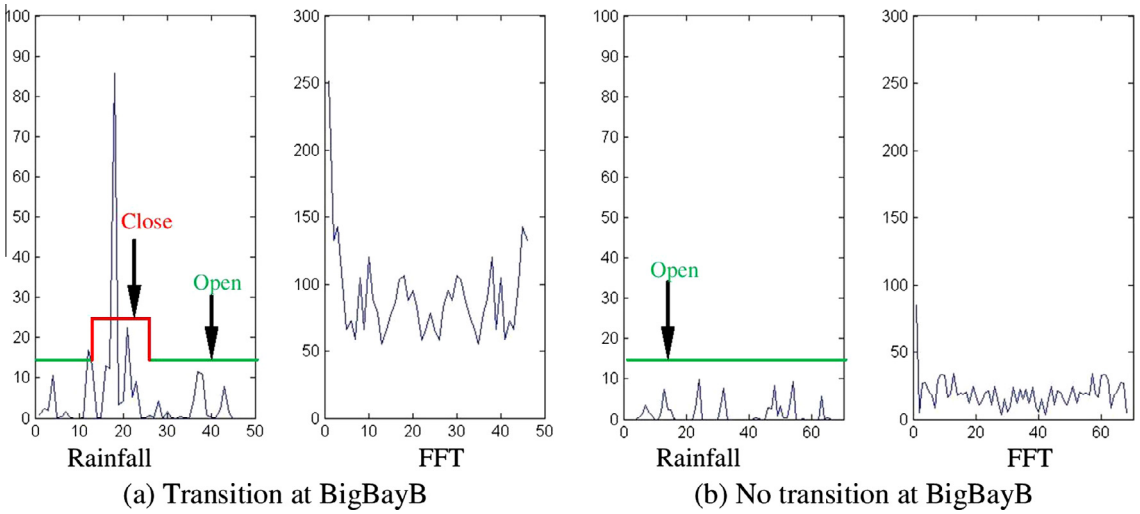


Fig. 4 – Performance of Fourier Transformation features on transitions.

Fig. 2(b) where no transition of states occurs whereas high frequency components are stronger in Fig. 2(a) where transition of states occurs. Also as the features are transformed, k-NN classifier does not suffer from the noises present in the raw data. Combined, k-NN with Fourier transform performs best of these data sets.

Table 7 – MCC scores obtained using Piecewise Aggregate Approximation Features. Bold indicates the best performance in a row.

	DT	SVM	MLP	BN	RBF	k-NN
Big Bay B	0.261	0.003	0.126	**0.288**	0.03	0.158
Big Bay C	0.238	−0.008	0.158	**0.269**	0.008	0.128
Duck Bay	0.273	−0.009	0.174	**0.337**	0.081	0.168
Dunnalley Bay A	0.174	0.105	0.094	**0.272**	0.059	−0.001
Hastings	0.145	0.043	0.139	**0.222**	0.053	0.13
Montagu	0.184	−0.001	0.118	**0.226**	0.063	0.101

Table 8 – MCC scores obtained using Wavelet Features. Bold indicates the best performance in a row.

	DT	SVM	MLP	BN	RBF	k-NN
Big Bay B	0.383	0.448	0.399	**0.533**	0.433	0.377
Big Bay C	0.335	0.422	0.333	**0.515**	0.409	0.351
Duck Bay	0.376	0.504	0.406	**0.559**	0.446	0.334
Dunnalley Bay A	0.331	0.267	0.214	**0.362**	0.33	0.117
Hastings	0.407	**0.543**	0.432	0.543	0.488	0.324
Montagu	0.491	**0.548**	0.416	0.543	0.497	0.374

Table 9 – Best performance and corresponding combinations of feature set and classifiers in each location.

	MCC	TP (Closure)	TN (Open)	Overall accuracy	Feature set	Classifier
Big Bay B	0.814	85.78	95.25	92.53	Fourier Transform	k-NN
Big Bay C	0.818	85.76	95.66	92.95	Fourier Transform	k-NN
Duck Bay	0.806	85.35	94.86	92.23	Fourier Transform	k-NN
Dunnalley Bay A	0.723	81.24	90.62	89.53	Fourier Transform	k-NN
Hastings	0.838	86.33	96.97	94.67	Fourier Transform	k-NN
Montagu	0.856	90.06	95.41	93.35	Fourier Transform	k-NN

The Bayesian Network is the overall winner using *Piecewise Aggregate Approximation Features* (Table 7) and this feature also performs worse than most of the other features. The Bayesian Network is also the overall winner using *Wavelet Features* (Table 8). The *Wavelet Feature* is the second best performer (the *Fourier Transform Feature* is the best one). This is once again because of the fact that transitions are associated with closure events and frequency domain features can capture that very well.

5.3. *Best combination*

The best performing combination of different features and classifiers along with MCC score, accuracy of 'Open' classification, accuracy of 'Close' classification, and overall accuracy are presented in Table 9. Note that the balancing method is 'over-sampling'. In all locations the best performing feature is Fourier Transformation Features and the best classifier is *k*-NN. Both 'Close' and 'Open' classes were recognized with high accuracy. Fourier Transform captures the time series in this case. The reason is that the transition state ('Close') is presented by high frequency components whereas a steady state, like 'Open', is represented mostly by low frequency components.

6. Conclusion

In this paper we have evaluated the effectiveness of a time series prediction framework to accurately predict shellfish farm closure using rainfall data. The data mining challenges underlying the research are: (1) the identification of features that best represent rainfall patterns leading to farm closure, and (2) the best ways to deal with the data imbalance that naturally occurs with such problems due to the infrequency of one event. Different combinations of feature extraction methods, class balancing algorithms, and classifiers were evaluated on six different locations in Tasmania. The class balancing method (over-sampling/under-sampling) improves recognition accuracy. The *Fourier Transformation Feature* combined with the *k*-NN classifier performed better than other combinations on this particular problem. In future we aim to undertake similar studies to find appropriate combinations in multivariate time series data.

REFERENCES

[1] Rahman A, Verma B. A novel layered clustering based approach for generating ensemble of classifiers. IEEE Trans Neural Networks 2011;22(5):781–92.

[2] Rahman A, Verma B. Ensemble classifier generation using non-uniform layered clustering and genetic algorithm. Elsevier Knowl Based Syst 2013;43:30–42.

[3] Rahman A, Shahriar MS. Algae bloom prediction through identification of influential environmental variables: a machine learning approach. Int J Comput Intell Appl 2013;12(2). http://dx.doi.org/10.1142/S1469026813500089.

[4] D'Este C, Rahman A, Turnbull A. Predicting shellfish farm closures with class balancing methods. Aust Artif Intell Conf 2012. LCNS 7691, 39–48.

[5] Rahman A, D'Este C, Timms G. Dealing with missing sensor values in predicting shellfish farm closure. In: Proc. of IEEE intelligent sensors, sensor networks and information processing (ISSNIP), Melbourne, Australia; 2013. p. 351–6.

[6] D'Este C, Rahman A. Similarity weighted ensembles for relocating models of rare events. Multiple Classifier Syst 2013. LCNS 7872, 25–36.

[7] Rahman A, D'Este C, McCulloch J. Ensemble feature ranking for shellfish farm closure cause identification. In: Proc. workshop on machine learning for sensory data analysis in conjunction with australian AI conference, Dunedin, New Zealand; 2013. DOI: http://dx.doi.org/10.1145/2542652.2542655.

[8] Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intell Data Anal 2002;6(5):429–50.

[9] Rahman A, Smith D, Timms G. A novel machine learning approach towards quality assessment of sensor data. IEEE Sens J 2014;14(4):1035–47.

[10] Yang Z, Gao D. An active under-sampling approach for imbalanced data classification. Proc. Fifth International Symposium on Computational Intelligence and Design (ISCID) 2012;2:270–273.

[11] Sugimura H, Matsumoto K. Classification system for time series data based on feature pattern extraction. IEEE International Conference on Systems, Man, and Cybernetics (SMC) 2011;1340–1345.

[12] Wu G, Yang J. A representation of time series based on implicit polynomial curve. Pattern Recognit Lett 2013;34(4):361–71.

[13] Xingye L, Tian T. Time series recognition based on wavelet transform and Fourier transform. IEEE Symp Ind Electron Appl 2010:722–6.

[14] Guo CH, Li HL, Pan DH. An improved piecewise aggregate approximation based on statistical features for time series mining. In: Bi YX, Williams MA, editors. Proc. international conference on knowledge science, engineering and management (KSEM). Berlin, Heidelberg: Springer-Verlag; 2010. p. 234–44.

[15] Burger J. A Basic Introduction To Neural Networks. Source: http://pages.cs.wisc.edu/~bolo/shipyard/neural/local.html. 2013.

[16] MathWorks. Support Vector Machines (SVM). Source: http://www.mathworks.com.au/help/toolbox/bioinfo/ug/bs3tbev-1.html. 2012.

[17] Raileanu EL, Stoffel K. Theoretical comparison between the gini index and information gain criteria. Annal Math Artif Intell 2004;41(1):77–93.

[18] Pavlovic VJ, Frey BS, Huang T. Time-series classification using mixed-state dynamic Bayesian networks. In: Proc. IEEE conference on computer vision and pattern recognition (CVPR), Fort Collins, USA, vol. 2; 1999. p. 2609–15.

[19] D' Este C, Rahman A, Turnbull A. Predicting shellfish farm closures with class balancing methods. Aust Artif Intell Conf 2012;7691:39–48.

[20] Queensland Government. SILO climate data. link: http://www.longpaddock.qld.gov.au/silo/. 2013.

[21] Australian Government. Bureau of Meteorology. link: http://www.bom.gov.au/. 2013.

[22] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. SIGKDD Explorations 2009;11(1):10–18.

[23] Rahman A, Murshed M. Feature weighting methods for abstract features applicable to motion based video indexing. IEEE Int Conf Inform Technol: Coding Comput (ITCC) 2004;1:676–80.

[24] Rahman A, Murshed M. Feature weighting and retrieval methods for dynamic texture motion features. Int J Comput Intell Syst 2009;2(1):27–38.

[25] Nguyen MH, Cooper EW, Kamei K. Borderline over-sampling for imbalanced data classification. Int J Knowl Eng Soft Data Paradigm 2011;3(1):4–21.

[26] Gu Q, Cai Z, Zhu L, Huang B. Data mining on imbalanced data sets. In: International conference on advanced computer theory and engineering, Phuket Thailand; 2008. p. 1020–4.

[27] Zhao M, Chai Q, Zhang S. A method of image feature extraction using wavelet transforms. Int Conf Emerg Intell Comput Technol Appl 2009:187–92.

[28] Tran DJM, Lim C, Abeynayake C, Jain L. Feature extraction and classification of metal detector signals using the wavelet transform and the fuzzy ARTMAP neural network. J Intell Fuzzy Syst 2010;21(1–2):89–99.

[29] Yip SC. DFT based feature extraction with non-uniform spectral compression for robust speech recognition. In: IEEE international conference on acoustics, speech, and signal processing, Orlando, USA, vol. 4; 2002.

[30] Wu YL, Agrawal D, Abbadi EA. A comparison of DFT and DWT based similarity search in timeseries databases. In: International conference on information and knowledge management, Washington, USA; 2000. p. 488–95.

[31] Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57.

[32] Liu A, Ghosh JE, Martin C. Generative oversampling for mining imbalanced datasets. DMIN 2007:66–72.

[33] Zielesny A. From Curve Fitting to Machine Learning: An Illustrative Guide to Scientific Data Analysis and Computational Intelligence. Springer; 2011. 53–147.